ACCURACY OF GENOMIC SELECTION: COMPARING THEORY AND RESULTS

B. J. Hayes[1], H. D. Daetwyler[2,3], P. Bowman[1], G. Moser[3], B. Tier[4], R. Crump[4], M. Khatkar[5], H. W. Raadsma[5], and M. E. Goddard[1,6]

[1]Biosciences Research Division, Department of Primary Industries Victoria, 1 Park Drive, Bundoora 3083, Australia, [2]The Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, EH25 9PS, United Kingdom, [3]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, The Netherlands, [5]Centre for Advanced Technologies in Animal Genetics and Reproduction (ReproGen), University of Sydney, Camden NSW 2570, Australia, [4]Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351, Australia, [6]Faculty of Land and Food Resources, University of Melbourne, Parkville 3010, Australia.

SUMMARY

Deterministic predictions of the accuracy of genomic breeding values in selection candidates with no phenotypes have been derived based on the heritability of the trait, number of phenotyped and genotyped animals in the reference population where the marker effects are estimated, the effective population size and the length of the genome. We assessed the value of these deterministic predictions given the results that have been achieved in Holstein and Jersey dairy cattle. We conclude that the deterministic predictions are useful guide for establishing the size of the reference populations which must be assembled in order to predict genomic breeding values at a desired level of accuracy in selection candidates.

INTRODUCTION

Genomic selection refers to the selection of animals for breeding based on genomic breeding values. Meuwissen . (2001) demonstrated using simulation that the accuracy of genomic breeding values can be very high if they are predicted from a large number of DNA markers. Provided the markers are dense enough, the accuracy of genomic breeding values will depend on the number of individuals genotyped and phenotyped in the reference population where the effect of the markers are predicted, the heritability of the trait, and the number of independent loci or chromosome segments in the population (Goddard 2008; Daetwyler *et al.* 2008). Goddard (2008) and Hayes *et al.* (2009) further derived deterministic predictions of the number of independent chromosome segments based on the effective population size and the length of the genome of the species in question. These deterministic predictions would have great value in guiding the design of experiments to implement genomic selection if the accuracy they predicted agreed with that observed in real data. Such data is now available; recently, Van Raden *et al.* (2009) reported accuracies of genomic breeding values as high as 0.75 for total merit index in Holstein Friesian dairy cattle using 38416 single nucleotide polymorphism (SNP) markers genotyped in 3576 progeny tested bulls. Accuracies of genomic selection are also available for Australian Holstein Friesian and Jersey cattle, using a similar number of SNPs.

The aim of this paper was to assess the value of the deterministic predictions of accuracy of genomic breeding values given results that have been achieved in Holstein and Jersey dairy cattle.

MATERIALS AND METHODS

In Daetwyler *et al.* (2008) the accuracy of genomic breeding values was predicted as $r = \sqrt{Nh^2/(Nh^2 + q)}$ where N = number of individuals genotyped and phenotyped in the reference population, h^2 = heritability of trait or reliability of breeding values in the reference population, q = number of independent chromosome segments in the population. Daetwyler et al.

(2008) also proposed a corrected for their prediction when $N \geq q$. The correction was to add $r^4 q/(2N)$ to the above prediction to get the final accuracy. As $N \geq q$ for most of the situations we will investigate, we will use the accuracy from the above equation with the correction.

In Goddard (2008), the accuracy of genomic breeding values was predicted as $r = \sqrt{\left[1 - \lambda/(2N\sqrt{a}) * \ln((1 + a + 2\sqrt{a})/(1 + a - 2\sqrt{a})\right]}$ where $a = 1+2\lambda/N$, and $\lambda = qk/h^2$, with $k = 1/log(2N_e)$, where N_e is the effective population size. Note that this derivation assumes that σ_e^2 is close to the phenotypic variance. For both predictions, the value of q used was the number of independent chromosome segments, $2N_eL$, where L is the length of the genome in Morgans (Hayes *et al.* 2009). The difference between the formula of Daetwyler *et al.* (2008) and Goddard (2008) potentially arises because Goddard (2008) assumed that that the effect estimate for common QTL is more accurate for QTL with intermediate allele frequency, because they explain more of the genetic variance than QTL with extreme allele frequency In contrast, Daetwyler *et al.* (2008) assumed the accuracy of estimating QTL effects was equal regardless of their frequency. The accuracy of genomic breeding values for the two deterministic predictions were compared for a range of heritabilities, N=5000 and N_e=100.

Accuracy of breeding values from the two predictions were also compared to accuracies of genomic breeding values reported by VanRaden *et al.* (2009) and United States Department of Agriculture results (http://aipl.arsusda.gov/reference/genomic_comparison_yng_0901.htm) for total net merit in Holstein Friesian cattle and Jersey cattle, and our own results in Australian data for these two breeds. The phenotypic records in the reference population were daughter yield deviations (DYD) for total merit index for the US data or de-regressed breeding values for Australian Profit Ranking (APR) in the Australian data. The average reliability of the DYD in the reference population was 0.9. In order to deterministically predict the accuracy that these experiments could have achieved, an assumption of the N_e in each breed was required. Young and Seykora (1996) gave an estimate of 100 for the N_e of US Holsteins. The N_e in Australian Holsteins is similar (deRoos *et al.* 2008). For US Jerseys, the effective population size is smaller, with a recent estimate of 30 (Weigel *et al.* 2008). The N_e of Jersey's in Australia is likely to be similar given the large contribution of US Jersey bulls to the Australian population. Given these estimates of N_e in the two breeds, we used N_e=100 in the predictions for Holsteins and 30 in Jersey's. A genome length of 30 Morgans was assumed.

RESULTS AND DISCUSSION

The accuracies of genomic breeding value predicted by Goddard (2008) and Daetwyler *et al.* (2008) are similar, though Daetwyler *et al.* (2008) would predict a lower accuracy of breeding value at low to moderate heritabilities given the same number of independent chromosome segments and number of phenotypic records, Figure 1. Both deterministic predictions agreed fairly well with the accuracies of genomic breeding value reported for US and Australian Holstein Friesian and Jersey dairy cattle, Figure 2. The % error was low for the Goddard prediction vs the US Holstein data at 3%.. However in the Australian Holstein data the observed accuracies were somewhat higher than the predictions. This may just reflect a small validation sample used in the Australian data leading to a large standard error for the estimate of reliability.

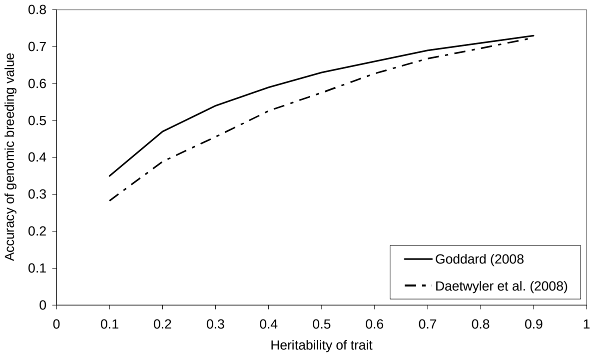

Figure 1. Accuracy of genomic breeding values with 5000 phenotypic records, effective population size of 100 and increasing heritability, predicted by the deterministic formula of Goddard (2008) or Daetwyler et al. (2008).

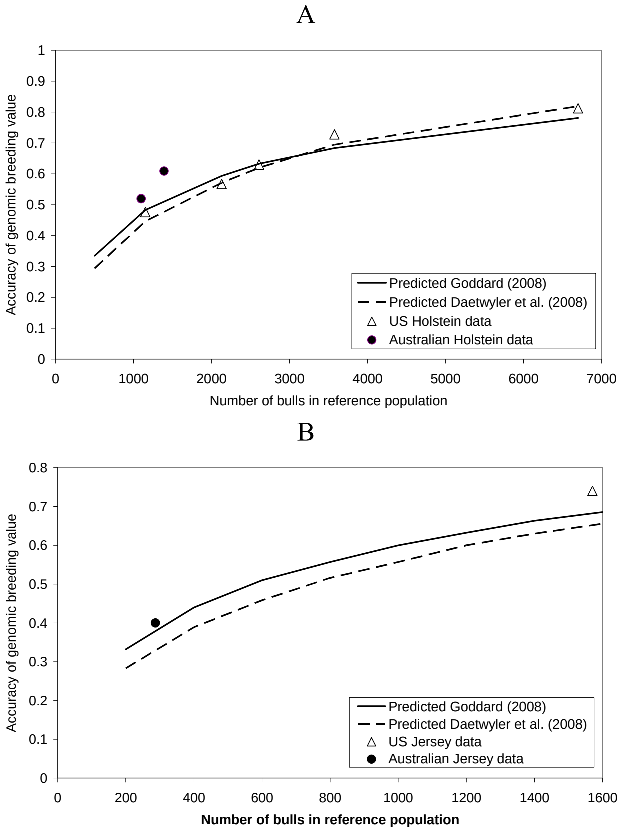

Figure 2. A. Accuracy of genomic breeding values from the deterministic prediction of Goddard (2008) and Daetwyler *et al.* (2008) with N_e=100, and accuracy of genomic breeding value for total merit index or Australian Profit ranking in US or Australian Holstein Friesian cattle. B. Accuracy of genomic breeding values from deterministic predictions with N_e=30, and accuracy of genomic breeding value for total merit index or Australian Profit ranking in US or Australian Jersey cattle respectively.

Another contributing factor may be that the deterministic predictions assume that the accuracy of breeding values is a result of the SNPs capturing the effect of QTL, whereas some of the accuracy of genomic breeding values in livestock populations may be a result of the SNPs capturing the effect of relationship, particularly if there are large half sib families in the population (eg. Habier *et al.* 2007). For comparison, the accuracy of parent average breeding values for net merit available for young bulls in the US data was 0.37 (VanRaden *et al.* 2009).

The deterministic method of Goddard (2008) used here assumes a normal distribution of QTL effects. For the majority of traits studied by Van Raden *et al.* (2008), methods for predicting genomic breeding values which assumed a normal distribution of quantitative trait loci (QTL) effects performed almost as well as methods assuming a exponential distribution of QTL effects. The exception was traits with a QTL of known large effect, eg. fat percentage (Grisart *et al*. 2003). For such traits, the deterministic prediction of Goddard (2008) would under-predict accuracy of genomic selection. The accuracies of prediction also depend on N_e. The values of N_e used here are estimates of N_e in the recent past, however N_e in cattle has been much larger historically. It is not clear how the change in historical N_e should affect accuracy of genomic breeding values. Nevertheless, using current N_e gave good agreement between predictions and observed results.

CONCLUSIONS

The deterministic predictions of accuracy of genomic selection presented by Goddard (2008) extended by Hayes *et al.* (2009), and that of Daetwyler *et al.* (2008) agree well with observed accuracies of genomic selection in US and Australian Holstein Friesians and Jerseys. We can conclude that these deterministic predictions are a useful tool to guide design of genomic selection experiments, for example how large should the reference population be to achieve a desired level of accuracy. It must be noted we have compared predicted and observed accuracies of genomic breeding value for a situation where phenotypes were very accurate predictors of breeding value. The performance of the deterministic predictions of both Daetwyler *et al.* (2008) and Goddard (2008) should be also evaluated in other situations where the heritability of the trait is lower, as the difference predicted accuracy of genomic selection is greater at lower heritabilities.

ACKNOWLEDGMENTS

The authors are grateful to Curt van Tassell and Tad Sonstegard from the USDA for providing genotypes under a collaborative agreement between USDA and Department of Primary Industries Victoria. We are also grateful to Paul Van Raden and George Wiggans of USDA for providing some of the information required to assemble Figure 2.

REFERENCES

Daetwyler, H.D, Villanueva, B. and Woolliams, J.A. 2008. *PLoS ONE* **3**: e3395.
de Roos, A.P., Hayes, B.J., Spelman, R.J. and Goddard, M.E. 2008 *Genetics* **179:**1503
Goddard, M.E. 2008. *Genetica*. doi: 10.1007/s10709-008-9308-0.
Habier D, Fernando RL, Dekkers JC. 2007. *Genetics.* **177**:2389
Hayes, B.J., Visscher, P.M., and Goddard, M.E. 2009. *Genet. Res.* **91**:47
Meuwissen, T.H.E., Hayes, B.J, and Goddard, M.E. 2001. *Genetics* **157**: 1819
VanRaden, P.M. *et al.* 2009 *J. Dairy Sci.* **92**:16
Weigel, K. A. *J.Dairy.Sci.***84**(E. Suppl.):E177
Young, C.W. and Seykora, A.J. 1996.. *J. Dairy Sci.* **79:**502